基于加权回归分析以支持向量机为风险控制的多因子选股模型实证研究

作者：李皓辰，闫昱 通讯作者：经玲，王燕飞
（中国农业大学理学院，北京市，邮编：100083）

摘 要：考虑股票选取的多因子问题，在传统模型的基础上，利用MATLAB软件建立使用加权回归（权由日期和涨跌幅综合决定）的股票基本面指标、技术指标对相对收益率的多因子模型，并且引入支持向量机作为风控，最终获得了一个收益较好的量化投资模型。

0 引 言

如何科学地进行投资是市场经久不衰的研究话题，而随着量化投资的概念在世界范围内兴起，各种量化投资的方法也逐渐受到中国投资者的关注，另一方面，量化投资的思想也逐渐为投资者所接受，开始成为证券界的研究热点。

本文将以研究热点多因子模型为基础使用统计学中的加权回归方法对模型进行了改良，之后又使用了支持向量机的算法来对风险进行控制。

1 综 述

多因子模型最早来源于詹森提出的Alpha收益概念，这一概念来自于詹森对CAPM模型的扩充模型如下:

（1）

其中，代表着资产的实际收益率；代表市场组合预期收益率；

之后在1993年Fama和French提出了Fama-French选股模型，这一模型表达式如下：

（2）

这一模型可以看作目前广泛使用的多因子模型的雏形，其中为无风险收益率。SMB为规模市场中因子组合收益率，常用大小市值股票组合收益率之差表示。HML为公司价值因素，常使用高低账面股票市值比股票组合的收益率之差来表示。其中反映了这个投资组合对规模因子的敏感度，而表示了投资组合对价值因子的敏感度，其中的表示了投资组合对价值因子的敏感度。

经过国内学术界学者的实证研究，FF模型在国内市场被认为是有效的，也就是说中国股市中还是存在一定的估值效应和规模效应的，但是FF模型中在实际应用也有自己的不足，并且其并没有发展成完整的理论体系来对市场中所有股票的情况加以解释，并且三因子模型只能描述投资标的中较小的一部分特征，在实际投资领域还有这更多的基本面和技术变得指标被投资者所使用和信赖。我国的学者也做了许多相关的基于A股市场的验证，在传统理论中，由于投资者交易理念、信息环境等因素的区别，基于美国成熟的股票市场的数

据为样本进行分析的资产定价模型往往与新兴资本市场有所不同。近期，在参考文献[2]中，田利辉等人研究了五因子模型在中国证券市场中的表现，并且发现我国的定价因素与美国市场经验有较大区别。在参考文献[1]中，赵胜民等人研究了 Fama，French 在 2015 提出的五因子模型在中国 A 股市场中的表现，得出了虽然具有一定的有效性但是表现却不如三因子模型的结论。

为了对 FF 模型进行补充，出现了多因子选股模型，其基本假定就是股票的未来走势将是其历史的重演，并且股票的收益率受到上市公司财务指标及一些行情指标的影响，多因子选股模型就是一句一定的条件，筛选出可以战胜市场的投资组合。其模型如下：

在实践之中，往往使用多元线性回归的方法对进行求解，往往使用最小二乘法即等价于求解如下问题：

我们提出了在回归时使用加权回归的方式进行求解，那么问题转化为：

在模型之中相当于对模型的惩罚因子进行加权，即不是把每一天看的同等重要，只是在计算时要对 w(j)同时进行考虑。股票市场是经济的晴雨表，其有一定的周期性特点，所以在回归的过程中首先就是距离预测时间较近的日期应该予以较大的权重，并且考虑的股票的周期性，如果取周期是两年的话，那么相应的两年前的数据自然应该比一年前的数据取更大的权重。而且对于涨跌幅来说惩罚系数也应有相应的考量，因为在波动不那么剧烈的时候，即使预测出现偏差也不会带来较大的亏损，但是如果错过了较大的涨幅或者没有避开较大的跌幅那么就会对收益产生巨大的影响，所以权重过于涨跌幅也应该是正相关。

这是对多因子模型的主体的计算，另一方面，在量化投资的过程中最重要的一点是对风险的控制，即对回撤的限制。所以在本文之中使用支持向量机对风险进行控制。

支持向量机(Support Vector Machine，SVM)是 Corinna Cortes 和 Vapnik 等人于 1995 年首先提出，它在解决小样本、非线性及高维模式识别中表现出许多独特的优势，并能够推广应用到函数拟合等其他机器学习问题中。

在机器学习中，支持向量机是与相关的学习算法有关的监督学习模型，可以分析数据，识别模式，用于分类和回归分析。在股票预测中，可以将股票的涨和跌看做两种状态，在这种情况下就可以将原本的连续性变量：涨跌幅变为离散型的数据。

在过去的研究之中，往往使用支持向量机对大盘指数进行预测，也取得了不错的效果

在过去学者们的研究中表示，支持向量机对股票跌幅的预测往往很准确，而多因子模型在回归的过程中往往难以对跌幅方面进行很好的防范，这就导致了在具体的实验之中虽然获得了很好的收益率但往往伴随着很大的风险，引入支持向量机作为风险控制，当支持向量机预测结果表示未来股票将会下跌时，则即使多因子模型提示其预测涨幅较大也不选择买入，则可获得平稳的超额收益。

本文的主要贡献在于：本文建立了一个全新而有效的量化投资模型并且将新的回归模

式引入了多因子模型之中。

2 模 型 建 立

2.1 传统因子模型

2.1.1 传统三因子模型

传统的三因子模型是 Fama 和 French 在 1993 年提出的。模型认为，一个投资组合(包括单个股票)的超额回报率可由它对三个因子的表现来解释，这三个因子是：市场资产组合、市值因子、账面市值比因子。这个多因子均衡定价模型可以表示为：

其中表示时间 t 的无风险收益率；表示时间 t 的市场收益率；表示资产 i 在时间 t 的收益率；是市场风险溢价，为时间 t 的市值(Size)因子的模拟组合收益率，为时间 t 的账面市值比因子的模拟组合收益率。

分别是三个因子的系数，采用多元回归的方式来计算，回归模型表示如下：

值得一提的是，表示的是三因子模型里面尚未解释的超额收益。若是三因子模型已经可以完全解释各种风险带来的超额收益，那么任何一个股票以及任何一个投资组合的的真实值应该为 0。

2.1.2 三因子模型基础上的五因子模型

三因子模型在后续的研究中，学者发现许多股票的值显著不为 0，于是在三因子模型的基础上，产生的五因子模型，回归模型可表达如下：

其中，五因子模型比三因子模型里面多出来了两项：是高/低盈利股票投资组合的回报之差，而则是低/高再投资比例公司股票投资组合的回报之差。同样，表示的是五因子模型里面尚未解释的超额收益。

2.2 多因子模型

2.2.1 多因子模型简介

多因子选股模型是现在应用比较广泛的一种选股模型。与之前的模型不同，它进一步弱化了人为选择影响因素的步骤。基本原理就是采用一系列的因子作为选股标准，满足这些因子的股票则被买入，不满足的则卖出。

通常来说，多因子选股模型有两种判断方法，一是打分法，二是回归法。我们采用的是线性回归的方法。并在此基础上，创新性地使用了加权线性拟合的方法。

2.2.2 使用支持向量机

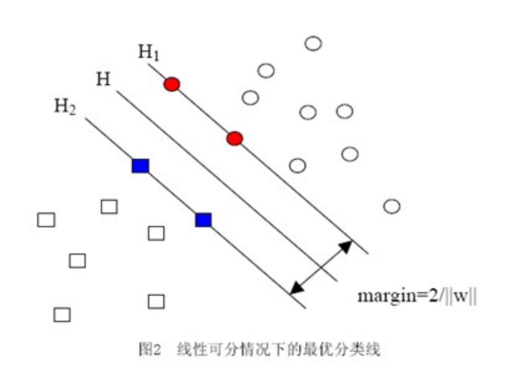


图2 线性可分情况下的最优分类线

支持向量机方法(SVM)就是从解决线性可分情况的最优分类出发的，其思想就是选取使“间隔”达到最大的那个法方向，相应得到的两条极端直线就是最优分类线，这是 SVM 的核心思想之一。具体来说就是，对于选定的法方向 w， 会有两条极端的直线，选取 b 使得要找的直线为两条极端直线“中间”的那条直线。我们使用支持向量机的时候将所有交易日看做两类：涨和跌，在此基础上我们始终支持向量机进行分类。

2.2.3 加权最小二乘线性拟合

一般曲线回归的方程为:

采用最小二乘法可以给出参数向量的估计,即的估计值应使下式

最小。 但是在实际应用中,一般的曲线方程 f(x,β)的形式都比较复杂,采用非线性优化技术求解上式时，对初始点具有较强的依赖性，很容易陷入局部最优解,从而给估计带来困难。为了解决上述问题,可以采用加权的最小二乘线性拟合：

即如果存在变换这样就将转化为了此时线性回归方程可以表示为：

通过最小二乘法对上式进行估计后再通过相应的逆变换得到原式中的估计。

3 实 证 研 究

在本文之中，使用的所有数据均来自于 wind 软件（wind 是一款可以通过 matlab 对接的十分有效的金融资讯平台），股票样本主要选取 2014 年 6 月 27 日到 2016 年 6 月 27 日的时间序列数据，并且选取的是上海交易所上市的公司。

我们选取了如下的十七个指标作为回归数据的参考，并且对相关系数做出了如下计算:

表 1 指标和其相关系数

特征名称	相关系数	特征名称	相关系数
前收盘价	-0.04404	均价	0.009259
开盘价	-0.04263	换手率	0.080641
最高价	-6.95E-05	持仓量	0
最低价	0.001029	持仓量变化	0.000912
收盘价	0.045941	相对发行涨跌幅	0.038018
成交量	0.074991	净流入资金	0.471579
成交额	0.056889	市盈率	0
涨跌幅	0.83079	市净率	0
振幅	0.047996		

在计算之后，我们选取相关系数最大的十个指标作为备选指标，之后采用多因子模型和加权多因子模型对全市场 1081 支股票分别进行拟合，下图给出中国石油、苏宁云商、大智慧、兴业银行、苏宁云商和蒙草抗旱的预测曲线在最近四百多天的价格的对比图，可以看出进行加权之后多因子模型的预测精度并没有很大下降，但是对于大涨大跌时的判断准确了许多。

之后我们给出中国石油部分指标和部分预测值：

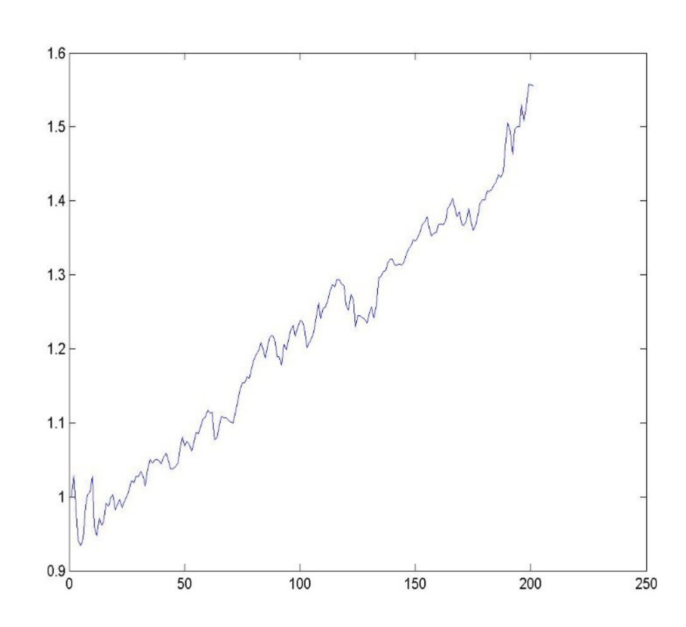
表 2 中国石油指标和预测值

日期	开盘价	最高价	收盘价	交易量(股)	涨跌幅	加权预测收盘价	预测收盘价
2016/5/27	7.21	7.22	7.2	9075211	-0.01	7.210832695	7.2730830
2016/5/26	7.21	7.25	7.22	14301584	0.01	7.232441659	7.303096346
2016/5/25	7.22	7.24	7.19	11116979	-0.03	7.20249207	7.259018861
2016/5/24	7.19	7.22	7.22	14916983	0.03	7.23137726	7.304104514
2016/5/23	7.21	7.23	7.2	10559126	-0.01	7.211223125	7.272259365
2016/5/20	7.17	7.21	7.21	8715582	0.04	7.218397176	7.295062182
2016/5/19	7.18	7.23	7.18	16951884	0	7.193572545	7.258825453
2016/5/18	7.21	7.22	7.2	28808330	-0.01	7.219066752	7.279707528
2016/5/17	7.21	7.25	7.24	17886829	0.03	7.252208087	7.322488281
2016/5/16	7.17	7.2	7.2	13220329	0.03	7.210854445	7.284880385
2016/5/13	7.2	7.25	7.19	21040632	-0.01	7.205722074	7.267961657
2016/5/12	7.21	7.21	7.2	17221805	-0.01	7.214203599	7.275184283
2016/5/11	7.21	7.25	7.22	14851902	0.01	7.232249578	7.300071783
2016/5/10	7.18	7.22	7.19	13713153	0.01	7.20166425	7.268748139
2016/5/9	7.32	7.32	7.2	34754111	-0.12	7.22686855	7.259769811
2016/5/6	7.46	7.46	7.34	32726175	-0.12	7.365928828	7.398927185
2016/5/5	7.45	7.48	7.46	17745527	0.01	7.47342592	7.541216781
2016/5/4	7.48	7.51	7.46	21305853	-0.02	7.476426874	7.537052938
2016/5/3	7.4	7.51	7.51	29245563	0.11	7.523621067	7.620060813

之后我们将支持向量机引入计量之中，在实验过程中我们使用了多种核函数进行计算，最终发现二次核函数的效果最佳。在训练时，将涨的天的指标,记

为 1，将涨的天的指标,记为 0，使用加权回归对模型进行求解，得到每一个的

股票的预测值为：。之后对整个市场中的所有股票的预测值进行排序，取前组合数的股票进行买入。



这是从 2016 年 6 月 27 日往前 200 天同时持有 100 值股票的带有对冲的回

测数据，从收益上来看获得了接近百分之六十的收益，如果能够保持年化收益

应该可以达到 75%以上。并且这一次回测是以散户为基准的，所以每天都有千

分之一的交易成本，如果是机构投资者这一成本将会大大降低可以获得更高的收益。

在交易过程中，在开始的二十天左右曲线具有较大的起伏，这正好与 2015 年下半年的震荡相吻合，而在 120 天左右的回撤与 A 股市场在 2016 年初的波动有关，所以虽然采用了对冲的策略，但是在具体的投资过程中往往没法应对股灾是情境。

为了避免实验的偶然性，我们还对四组数据每组数据试验了 60 支到 150 支股票，并且每一组收益都在百分之五十以上。

4 结　论

股票的预测理论一直是学者们所关注的热点问题，而从传统统计学、经济学和机器学习等学科都可以提出种类繁多的预测理论，本文所探讨的就是利用一种结合加权线性回归和支持向量机进行预测的准确性和利用此模型进行投资的有效性。

在本文之中，首先我们从投资时风险大小的计量出发，考虑对模型的惩罚应基于亏损的大小，因此使用了加权线性回归对经典的多因子模型进行了改进，并且在对权重进行合理考虑后发现虽然表面上看回归精度有所下降，但是却大大提高了收益，同时另一方面在加入了支持向量机对股票的涨跌进行二分类预测后起到了风控的作用。

最终从实用的角度出发我们获得了拥有较高收益量化模型，；该模型指导下的投资不仅收益可观，而且具有很好的稳定性，适合风险偏好型投资者采纳。同时佐证了中国证券市场具有弱有效性。不仅对现实之中的优质选股问题提供

了一定的方案，同时该问题还可用到公司财务评估、水质污染计量和土地资源评估等各个领域。

5 参 考 文 献

[1] 赵胜民，闫红蕾，张凯 Fama-French 五因子模型比三因子模型更胜一筹吗——来自中国 A 股市场的经验证据 [j].南开经济研究，2016（2）：41-59

[2] 田利辉，王冠英。我国股票定价五因素模型：交易量如何影响股票收益率？ [j].南开经济研究，2014（2）：54-75

[3] 张翔宇，王富森。基于支持向量机的上证指数预测研究 [j].商业经济，2011（3）：104-106.

[4] 黄宏运，王梅，朱家明。基于多元回归分析的多因子选股模型 [j].通化师范学院学报（自然科学）2016（4）：44-46

[5] 陈青，李子白。中国流动性调整下的 CAPM 研究 [j].数量经济技术研究，2008（6）：66-78

[6] 王维贤，陈利军。股票价格预测的建模与仿真研究[j].计算机仿真，2012（1）：344-347

[7] 顾亚祥，丁世飞。支持向量机研究进展 [j].计算机科学，2011，38（2）：14-18

[8] 黄朋朋，韩伟力。基于支持向量机的股价反转点预测 [j].通计算机系统应用，2010（9）：214-218

[9] 张玉川，张作泉。支持向量机在股票价格预测中的应用 [j].北京交通大学学报，2007（6）：73-76

[10] 吴世农，许年行。资产的理性定价模型和非理性定价模型的比较研究——基于中国股市的实证研究 [j].经济研究，2004（6）：105-116

[11] Cristianini N,Taylor JS 。An introduction to support vector machines and other kernel-based learning methods [A].Electronic Industry Press,Beijing，[C]2004

[12] Fama E.F. ， French K.R.A Five-Factor Asset Pricing Model [j]. 通 Journal of Financial Economics，2015,116（1）：1-22

[13] J.Patel， S.Shah， P.Thakkar and K.Kotecha.Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques [J]. Expert Systems with Applications. 2015, 42(1):259–268.